

FineTuneLab System Overview

November 30, 2025 - December 14, 2025

System Overview

Welcome to FineTuneLab

HIGHLIGHTS

- + FineTuneLab is an AI operations platform for managing LLM interactions, fine-tuning models, and monitoring performance.
- + This system has processed 82 conversations and 487 messages in the selected time period.
- + Key capabilities include: conversation management, model fine-tuning, analytics dashboards, and automated quality evaluation.
- + This report shows current system health and explains the metrics you'll be working with.

Key Metrics Explained

Metric	What It Measures	Why It Matters
Success Rate	Percentage of requests completed	Indicates system reliability - higher is
Error Rate	Percentage of requests that failed	Inverse of success rate - lower is better
Latency	Time from request to complete response	User experience - faster is better
P95 Latency	95% of requests complete within this	Worst-case user experience indicator
Tokens	Units of text processed by the AI model	Directly impacts cost - 1 token "H 4
Cost per Message	Average spend per AI interaction	Budget efficiency metric
Quality Rating	User feedback score (1-5 scale)	Measures output quality and usefulness
Tool Success Rate	Percentage of tool calls that worked	Reliability of integrations and actions

These are the core metrics you'll see throughout the platform.

Current Baselines

DAILY CONVERSATIONS 12	SUCCESS RATE 97.5%	AVERAGE LATENCY 2.2s
P95 LATENCY 2.5s	QUALITY RATING 4.3/5.0	COST PER MESSAGE \$0.0880

Healthy System Indicators

Metric	Target Range	Current Value	Status
Success Rate	"e 95% (excellent: "e 98%)	97.5%	Good
Average Latency	< 3s (excellent: < 2s)	2.17s	Good
Quality Rating	"e 4.0 (excellent: "e 4.5)	4.3/5.0	Good
Error Rate	< 2% (excellent: < 1%)	2.50%	Needs Attention

These targets help you quickly assess system health.

[Available Tools](#)

Tool Name	Description	Usage Count	Success Rate
web_search	Search the web for	3	67%
dataset_manager	Manage training datasets	2	50%
analytics_export	Export analytics reports	1	100%
training_control	Control training jobs	1	100%
token_analyzer	Analyze token usage	1	100%

5 tools available in the system.

[Common Questions](#)

Question	Answer
What does "success rate" measure?	The percentage of AI requests that complete without errors.
What is a "token"?	A token is roughly 4 characters of text. Both input (your message) and output (AI response) consume tokens, which
Why does latency matter?	Latency is how long users wait for responses. Current
How is quality measured?	Users rate responses 1-5. We also track success/failure.
What should I monitor?	Watch the dashboard for error rate spikes (> 5%), latency
How much does this cost?	Current cost is \$0.0880 per message. Total spend this period:
Where do I find detailed logs?	Check the Analytics Dashboard for real-time data, or export
Who do I contact for help?	Reach out to the platform team on Slack (#finetunelab-

For more questions, check the internal wiki or ask in #finetunelab-support.